Gene Deletion Algorithms for Minimum Reaction Network Design by Mixed-integer Linear Programming for Metabolite Production in Constraint-based Models: gDel_minRN

Takeyuki Tamura (tamura@kuicr.kyoto-u.ac.jp)
Kyoto University
Ai Muto-Fujita
RIKEN Center for Biosystems Dynamics Research
Yukako Tohsato
Ritsumeikan University
Tomoyuki Kosaka
Yamaguchi University

Research Article

Keywords: algorithm, metabolic network, constraint-based model, flux balance analysis, mixed-integer linear programming, growth coupling, gene deletion

Posted Date: July 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-611371/v1

RESEARCH

Gene deletion algorithms for minimum reaction network design by mixed-integer linear programming for metabolite production in constraint-based models: gDel_minRN

Takeyuki Tamura[1*], Ai Muto-Fujita[2], Yukako Tohsato[3] and Tomoyuki Kosaka[4,5]

Correspondence:
amura@kuicr.kyoto-u.ac.jp
Bioinformatics Center, Institute
or Chemical Research, Kyoto
Jniversity, Uji, Kyoto, Japan
ull list of author information is
vailable at the end of the article

Abstract

Background: Genome-scale constraint-based metabolic networks play an important role in the simulation of growth coupling, which means that cell growth and target metabolite production are simultaneously achieved. To achieve growth coupling, a minimal reaction-network-based design is known to be effective. However, the obtained reaction networks often fail to be realized by gene deletions due to conflicts with gene-protein-reaction relations.

Results: Here, we developed gDel_minRN that determines gene deletion strategies using mixed-integer linear programming to achieve growth coupling by repressing the maximum number of reactions via gene-protein-reaction relations. Computational experiments were conducted in which gDel_minRN was applied to iML1515, a genome-scale model of *Escherichia coli*. The target metabolites were three vitamins that are highly valuable and require cost-effective bioprocesses for economics and the environment. gDel_minRN successfully calculated gene deletion strategies that achieve growth coupling for the production of biotin (vitamin B7), riboflavin (vitamin B2), and pantothenate (vitaminB5).

Conclusion: Since gDel_minRN calculates a constraint-based model of the minimum number of gene-associated reactions without conflict with gene-protein-reaction relations, it helps biological analysis of the core parts essential for growth coupling for each target metabolite. The source codes are implemented in MATLAB, CPLEX, and COBRA Toolbox. The obtained data and source codes are available on {{https://github.com/taketam/gDel-minRN}}

Keywords: algorithm; metabolic network; constraint-based model; flux balance analysis; mixed-integer linear programming; growth coupling; gene deletion

Background

Computational approaches are becoming increasingly important in the production of useful metabolites using microorganisms [1, 2, 3, 4, 5, 6, 7]. One of the most popular mathematical models in genome-scale metabolic engineering simulations is the constraint-based model. Constraint-based models mainly consist of metabolic networks and gene-protein-reaction (GPR) networks.

Metabolic networks represent the relationship between chemical reactions and compounds in cells. Many chemical reactions are catalyzed by enzymatic proteins encoded by genes. Therefore, metabolic networks can be controlled by gene dele-

tions through reaction deletions. The relationships between reactions and genes are represented by GPR networks, in which the relationships between genes, proteins, and reactions are represented by Boolean functions.

In the metabolic network part of the constraint-based model, steady states are assumed in which each metabolic reaction rate (flux) is constant. Such an analysis is called flux balance analysis (FBA) [8]. In FBA, (1) for each compound, the sum of the producing fluxes is equal to the sum of the consuming fluxes; (2) in each reaction, the fluxes of substrates and products must satisfy the ratio in the chemical reaction equation, and (3) the upper and lower bounds are given for each flux.

The constraint-based model includes a virtual reaction that represents the cell growth. The cell growth reaction of the constraint-based model was designed to match the results of the biological experiments. In the most standard FBA with constraint-based models, cell growth is maximized in the simulation because genotypes that result in higher cell growth are more likely to remain in the culture after repeated passaging. The cell growth reaction rate and the target metabolite-producing reaction rate are called the **growth rate (GR)** and **production rate (PR)**, respectively.

Therefore, in the simulation of useful metabolite production by FBA, we often evaluate PR when GR is maximized. When cell growth and the target metabolite production co-occur, we say that **growth coupling** is achieved. However, the number of metabolites for which growth coupling is achieved under natural conditions is limited. Therefore, it is often necessary to calculate the gene deletion strategy for the given constraint-based model and the target metabolite. (See also Sub figure 1A.)

Among the many existing methods[9, 10, 11, 12, 13, 14, 15, 16], one of the most efficient methods for calculating reaction deletion strategies for growth coupling is the elementary flux vector-based method [17]. The elementary flux vector-based method determines a non-decomposable flux distribution that includes the cell growth reaction and the target metabolite production reaction and deletes the reactions that are not used by the flux distribution. In other words, this method selects a minimal number of reactions to be used in the flow where cell growth forces the production of the target metabolite and deletes reactions that are not used. It was shown that the elementary flux vector-based method can compute the reaction deletion strate-

Figure 1 (A) Problem setting of this study. PR of the target metabolite is evaluated when the GR is maximized. (B) The idea of gDel_minRN algorithm. The maximum number of reactions are repressed via gene deletions for the growth coupling.

gies for growth coupling for most target metabolites for *E. coli* and *S. cerevisiae* under aerobic conditions by the combination of such core flows [17]. However, fewer than 10% of the reaction deletion strategies were feasible as gene deletion strategies because of the gene conflicts when the GPR network was considered[18, 19].

Therefore, it would be desirable if such a method based on core flow detection for reaction deletion strategies could also detect gene deletion strategies. However, it is not straightforward to directly extend the calculation of reaction deletion strategies to the calculation of gene deletion strategies.

In this work, to achieve growth coupling by gene deletions, we have developed a mixed-integer linear programming (MILP)-based algorithm, **gDel_minRN**, to calculate the gene deletion strategies that inactivate as many reactions as possible that are not essential for growth coupling (See also Sub figure 1B). gDel_minRN calculates **g**ene **del**etion strategies that obtain the **min**imum **r**eaction **n**etwork for growth coupling.

We conducted computational experiments on iML1515, a genome-scale constraint-based model of *E. coli*, the most common model organism. The target metabolites were riboflavin (vitamin B2), pantothenate (vitamin B5), and biotin (vitamin B7). Vitamins have been industrially produced by chemical synthesis and biosynthesis. Considering the sustainability, biosynthesis is more promising than chemical synthesis, which produces pollutants, and improvement of vitamins biosynthesis is still needed because high productivity and cost savings are important factors [20]. The reduction of metabolic pathways leads to efficient vitamin production by reducing the amount of protein required. When gDel_minRN was applied, gene deletion strategies for growth coupling were successfully obtained for these three vitamins.

The gene deletion strategies obtained by gDel_minRN do not contradict the GPR network, and allow us to design metabolic networks that achieve growth coupling of these compounds by repressing the maximum number of reactions. Therefore, if we analyze the gene deletion strategies obtained by gDel_minRN, we may be able to clarify the biological significance of the core part required for growth coupling for

Figure 2 (A) A toy example of the constraint-based model. Circles and rectangles represent metabolites and reactions, respectively. Black and white rectangles are external and internal reactions. r_1, r_6, and r_7 are the substrate uptake, cell growth, and target metabolite production reactions. $[\alpha, \beta]$ represents the lower and upper bounds of the reaction rates. (B) The optimistic and pessimistic flux distributions from the viewpoints of PR for each gene deletion strategy when GR is maximized. Deleting g_3 achieves the growth coupling since PR≥PRLB and GR≥GRLB are satisfied even for the pessimistic case of PR.

the target compounds without contradicting the gene-protein-reaction relationships. We conducted biological analysis of the obtained gene deletion strategy for biotin growth coupling.

Results

Developed algorithm

The developed algorithm gDel_minRN searches, using MILP, the flux and corresponding gene deletions that satisfy

(1) GR and PR are above the given thresholds, GRLB and PRLB,

(2) The number of reactions repressed by gene deletions is maximum,

(3) GR is maximized where (2) has a higher priority than (3).

It should be noted that the GR and PR obtained above are not always realized when GR is maximized without PRLB. Therefore, gDel_minRN tests whether the obtained gene deletion strategy achieves growth coupling under the condition that GR is maximized without PRLB. In particular, gDel_minRN checks the lowest PR value when GR is maximized. If the obtained gene deletion strategy does not achieve growth coupling in this pessimistic case, then the gene deletion strategy is added to the prohibited list and another gene deletion strategy is searched in the same way by MILP.

For example, suppose that GRLB=PRLB=1 in Sub figure 2A. When GR is maximized under the conditions of GR≥1 and PR≥1, the flux distribution for each gene deletion strategy is summarized in Table 1(A). Because deleting g_1, g_2, or g_5 cannot satisfy GR≥GRLB or PR≥PRLB, the gene deletion strategy candidates that can satisfy (1) are limited to $\{g_3\}$, $\{g_4\}$ and $\{g_3, g_4\}$. The number of repressed reactions by deleting $\{g_3\}$, $\{g_4\}$ and $\{g_3, g_4\}$ are 1, 0 and 1, respectively, as shown in Table 1(A). Therefore, gDel_minRN first selects the deletion of $\{g_3\}$ or $\{g_4\}$. When the gene deletion strategy is not uniquely determined under the condition that the

Table 1 (A) The flux distribution for each gene deletion strategy when GR is maximized under the condition with GR≥1 and PR≥1. (B) The priority of each gene deletion candidate and resulting flux distribution.

(A)

Deletion	v_1	v_2	v_3	v_4	v_5	v_6	v_7	#repressed reactions
g_1	-	-	-	-	-	-	-	cannot achieve GR≥1
g_2	-	-	-	-	-	-	-	cannot achieve GR≥1
g_3	5	0	0	0	0	5	5	1
g_4	10	9	1	1	0	9	1	0
g_5	-	-	-	-	-	-	-	cannot achieve PR≥1
g_3, g_4	5	0	5	5	0	5	5	1

(B)

Deletion	priority	v_1	v_2	v_3	v_4	v_5	v_6	v_7
g_3	1	5	0	5	5	0	5	5
g_3, g_4	1	5	0	5	5	0	5	5
g_4	2	10	10	0	0	0	10	0

number of repressed reactions is maximized, gDel_minRN selects the gene deletion strategy whose GR is maximum among them. However, in this case, GR is 5 for both $\{g_3\}$ and $\{g_3, g_4\}$. Therefore, deleting $\{g_3\}$ and $\{g_3, g_4\}$ have the same priority. Regardless of whether $\{g_3\}$ or $\{g_3, g_4\}$ is selected, GR=PR=5 is obtained and growth coupling is achieved as shown in Table 1(B).

gDel_minRN stops if the candidate of the gene deletion strategy achieves growth coupling. If growth coupling is not achieved, the obtained gene deletion strategy is added to the prohibited list, and gDel_minRN searches for the next solution. If no solution is obtained after the designated number of iterations, *maxloop*, gDel_minRN stops. Although the example is simple for illustration, gDel_minRN can be applied to complex GPR rules that combine AND and OR function. An AND function $y = x_1 \wedge x_2 \wedge \cdots \wedge x_k$ is converted into the linear constraints $-x_1 - \ldots - x_k + ky \leq 0$ and $x_1 + \cdots + x_k - y \leq k-1$. An OR function $y = x_1 \vee x_2 \vee \cdots \vee x_k$ is converted into the linear constraint $x_1 + \cdots + x_k - ky \leq 0$ and $-x_1 - \cdots - x_k + y \leq 0$. gDel_minRN cannot be applied directly to the case where NOT functions are included, but many latest genome-scale models such as iML1515 do not include NOT functions.

In gDel_minRN, we use MILP with PRLB to obtain the candidate for gene deletion strategies and then test whether growth coupling is achieved during GR maximization without PRLB. The reason why gDel_minRN maximizes the number of repressed reactions is that the more similar the flux distributions are when using

PRLB and when not using PRLB, the higher the success rate of the algorithm. This is also the reason why the second optimization target is the maximization of GR.

Computational experiments

All procedures in the computational experiments were implemented on a CentOS 7 machine with an Intel Xeon Processor with 2.30 GHz 18C/36T, and 128 GB memory. This workstation had CPLEX 12.8, COBRA Toolbox 2021 [21], and MATLAB R2017b. An auxiliary exchange reaction was temporarily added to the model to simulate the target metabolite production.

In the computational experiments, three vitamins, pantothenate (vitamin B5), biotin (vitamin B7), and riboflavin (vitamin B2), were used as target metabolites. These three metabolites are highly valuable, but no effective biosynthesis methods have been established. We applied gDel_minRN for growth coupling of these three target metabolites to iML1515[22], which is one of the most recent genome-scale constraint-based models of *E. coli* and includes 1515 genes, 2712 reactions, and 1877 metabolites.

Table 2 summarizes the gene deletion strategies obtained using gDel_minRN. Because the number of repressed reactions is maximized in gDel_minRN, the average number of deleted genes, 960.33, was almost twice as large as the average number of remaining genes, 554.67. When the obtained gene deletion strategy was applied and GR was maximized, the PR ratio to the theoretical maximum (TMPR) was 0.7444, 0.1004, and 0.1702, respectively. GR ratio to the theoretical maximum (TMGR) were 0.2485, 0.1702, and 0.1434, respectively. Because the minimum required PR/TMPR and GR/TMGR were 0.1 in the experiments, we can say that the strategies for pantothenate and biotin worked well, but that for riboflavin was not sufficient. The maximum computation time was approximately 6 h, which is within the acceptable range for individual calculations, but may not be suitable for batch calculations for hundreds of target metabolites. The first candidate gene deletion strategy obtained for each target metabolite failed to achieve growth coupling and succeeded in the second or third attempt using the prohibited list. Therefore, we can say that the iterative search algorithm using the prohibited list worked effectively.

Table 2 Three vitamins used as the target metabolites and the summary of the obtained gene deletion strategies by gDel_minRN. An auxiliary exchange reaction was temporarily added to the model to simulate each target metabolite production.

Target	#used genes	PR/TMPR	GR/TMGR	time	loop	abbreviation
Pantothenate (vitamin B5)	562	0.7444	0.2485	4h40m43s	3	pnto__R_c
Biotin (vitamin B7)	538	0.1004	0.1702	6h20m26s	2	btn_c
Riboflavin (vitamin B2)	564	0.0437	0.1434	2h58m49s	2	ribflv_c

Discussion

Biological analysis for biotin production

One of the motivations for developing gDel_minRN was to calculate the core parts required for growth coupling and to biologically elucidate which features are necessary for growth coupling and which are not. Among the three gene deletion strategies obtained by gDel_minRN, the most genes were deleted in the case of biotin. Therefore, the obtained biotin production pathway was analyzed biologically using Escher [23] and KEGG Mapper [24] as follows.

In the obtained pathway for biotin production by gDel_minRN, it was observed that the pathways from acetyl-CoA to acetate were removed from the map. The acetyl-CoA obtained in glycolysis was consumed in the TCA cycle or converted to acetate, and was also used to generate malonyl-CoA. Since malonyl-CoA is located at the beginning of the biotin-generating pathway, we hypothesized that by inhibiting the conversion of acetyl-CoA to acetate, acetyl-CoA that could not be fully consumed by the TCA cycle was used for biotin generation via malonyl-CoA.

To test this hypothesis, we revived all eight deleted genes (b0871, b2296, b0968, b2297, b2458, b4069, b3588, b1241) located on the pathways from acetyl-CoA to acetate. As a result, GR = 0.3341 and PR = 0 were obtained. This reinforces the hypothesis that by removing the acetyl-CoA to the acetate pathway, the substate used for cell growth was replaced by biotin production via malonyl-CoA.

Since the existing basic strategy for improving biotin productivity using bacterial cells is the overexpression of rate-limiting enzymes, removal of negative regulators and addition of intermediates or precursors [25], complete optimization of the metabolic pathways by altering the whole genomic network has not been extensively tested. The constructed pathway for biotin synthesis from iML1515, a recent solid computational model for *E. coli* metabolism, with the lowest number of reactions by gDel_minRN in this study showed new possibilities for the *E. coli* metabolic pathway

that can be changed from the original genome. Although the constructed pathway is stoichiometrically reasonable because iML1515 has almost complete metabolic network [22], it is not clear whether it can be created in *E. coli* real cells. Therefore, we considered this pathway from a biological point of view. The constructed pathway from glucose to biotin can be separated into two phases, from glucose to malonyl acyl-carrier-protein (ACP) and malonyl-ACP to biotin, respectively (Sub figure 3A). For biotin production, S-adenosylmethionine (SAM) and L-alanine are required to synthesize and adjust the production ratio in the upper pathway to drive the lower pathway (Sub figure 3A). The reactions in the lower pathway were not so unique because almost one connected pathway from malonyl-ACP to biotin in *E. coli* [20]. On the other hand, the biological consideration of the upper pathway, glucose to malonyl-ACP, revealed three notable characteristics. The most interesting characteristic was that nicotinamide adenine dinucleotide (NAD) was not used throughout the reactions. This result probably came from the calculation conditions for growth coupling with the minimum medium and glucose as the sole carbon substrate because all amino acids and nucleotides are required for synthesis, and the enzyme responsible for these reactions utilizes nicotinamide adenine dinucleotide phosphate (NADP) mainly as an electron carrier. In addition, the strategy of gDel_minRN is to reduce the reactions as possible then if electron carrier NAD not used in the pathway the many metabolic reactions can be reduced. Although it is very outlandish, since the dependency of NAD for biological metabolism is come from enzyme specificity, if there is no NAD-dependent enzyme, and NADP can drive all related reactions, NAD is not essential. Therefore, it is not biologically impossible. The second characteristic is the requirement for aerobic metabolism. In these reactions, a high amount of NADPH was produced from glucose to ribose 5-phosphate pathway, and oxidation was performed in dihydroxyacetone phosphate to glycerol 3-phosphate by glycerol-3-phosphate dehydrogenase, and NADP could then be produced (Sub figure 3B). Countering, in the opposite direction from glycerol 3-phosphate to dihydroxyacetone phosphate utilizing ubiquinone-8 (UQ8) as an electron acceptor to produce $UQ8H_2$ (Sub figure 3B). In addition, the reactions for pyruvate to lactate and succinate to fumarate generate $UQ8H_2$. These reactions cause high accumulation of $UQ8H_2$; oxidation is required to proceed with the metabolic reaction accomplished by using oxygen as an electron acceptor on the

Figure 3 The constructed pathway for biotin production. (A) Overview of the biotin synthesis pathway from iML1515 classified into two pathways as upper and lower pathway. (B) Precise flow of upper pathway, from glucose to malonyl-ACP. The number indicated with each arrows shows the flux value of each reaction. The abbreviations are as follows; NADPH, Nicotinamide adenine dinucleotide phosphate reduced form; UQ8, ubiquinone-8; ACP, acyl carrier protein; PEP, phosphoenolpyruvate; OAA, oxaloacetate; PRPP, Phosphoribosyl diphosphate.

respiratory chain, which also causes adenosine triphosphate (ATP) production in respiratory chains. High ATP production also requires not only the lower pathway but also nucleotide and amino acid synthesis. Therefore, this pathway requires oxygen or a respiratory oxidative substrate. We did not investigate the effect of the presence of oxygen on the constructed pathway. Therefore, future experiments should consider how substrate and culture conditions affect this pathway. The third characteristic is that the intermediates of this pathway do not consider about the cytotoxicity. The upper pathway utilizing methylglyoxal as the intermediate from dihydroxyacetone phosphate to lactate (Sub figure 3B). The methylglyoxal utilizing pathway is known in 1,2-propanediol producing bacteria but it shows that high cytotoxicity [26]. This pathway is possible but has problems. Several microorganisms for 1,2-propanediol production consider the pathway to not be exchanged because of the reduction in growth or production by the pathway [27]. This suggests that if we try to resolve more cell suitable pathways, we need some trick to avoid using the pathways from the literature to produce more realistic computational minimum pathway prediction for production. Finally, several problematic points for the construction or reproduction of this pathway in *E. coli* were found, but the constructed pathway was almost biologically possible in our consideration. Interestingly, when using a short and small number of reactions for some material production, cells can reduce the protein amount, which finally guides more efficient material production by the cell. The biological consideration of this pathway is only a knowledge base, and an experimental demonstration of this pathway on a cell should be performed in the future. To accomplish this, we need additional strategies, such as reduction of gene numbers for disruption or high number or gene disruption methods at the genomic scale. Alternatively, the use of semi-synthetic minimal cells is recommended to prove this pathway.

Comparison with existing computational methods

In the calculation of gene deletion strategies for growth coupling, it has been necessary to minimize the number of genes to be deleted in terms of cost and accuracy [28, 19]. However, gDel_minRN maximizes the number of reactions that are repressed to obtain the core part necessary for growth coupling, so it would rather delete as many genes as possible. Therefore, the obtained gene deletion strategies are quite different from those obtained using existing methods. Such a gene deletion strategy is helpful for biological analysis of which part of the constraint-based model is necessary for growth coupling but may not be practical for metabolite production with current metabolic engineering technology. However, it could be useful if zero-based DNA synthesis for metabolite production is possible in the future. In addition, under conditions where the product is obtained with growth coupling, it simplifies the actual production process and enables simultaneous production and cell maintenance in continuous culture.

On the other hand, for reaction deletions, the idea of finding a core network for growth coupling has been studied using elementary flux vector-based methods [17]. However, because the obtained reaction deletion strategies often conflict with GPR networks, it is difficult to extend the reaction deletion strategies to gene deletion strategies [19].

A number of mixed-integer linear programming (MILP)-based methods have been proposed for calculating gene or reaction deletion strategies that result in growth coupling [1, 2, 29, 30]. Solving MILP is an NP-complete problem and requires computation time proportional to the exponential function of the number of reactions. Many methods that are not limited to MILP have been proposed to speed up the computation time by avoiding the optimization of PR [9, 10, 11, 12, 13, 14, 15, 16]. However, to the best of our knowledge, there is no method for calculating the gene deletion strategy that results in a minimal network for growth coupling. Therefore, it is difficult to directly compare the performance of gDel_minRN with those of other methods in computational experiments.

Conclusion

In this study, we developed gDel_minRN to calculate gene deletion strategies that repress as many reactions as possible to achieve growth coupling. Computer ex-

periments using three vitamins as target compounds showed that we could find strategies that deleted more than 60% of all genes. Among them, we biologically analyzed the gene deletion strategy for biotin production and tested the hypothesis that deletion of genes in the pathway from acetyl-CoA to acetate replaces substrate consumption for cell growth with biotin production. Unlike existing biosynthetic methods for biotin production, the strategy obtained by gDel_minRN is based on a fundamental modification of the metabolic pathway. Existing computational methods aim to delete a small number of genes or compute core networks by deleting reactions, and their purpose is fundamentally different from that of gDel_minRN, which calculates core networks by gene deletion. Analyzing gene deletion strategies obtained by gDel_minRN is helpful for biological analysis for which parts are necessary for growth coupling.

Methods

Definition

Let $C = (M, R, S, L, U, G, F, P)$ be a constraint-based model, where $M = \{m_1, \ldots, m_a\}$, $R = \{r_1, \ldots, r_b\}$, $G = (g_1, \ldots, g_c)$, $F = (f_1, \ldots, f_b)$, and $P = (p_1, \ldots, p_b)$ are sets of metabolites, reactions, genes, Boolean functions, and the outputs of F, respectively. R always includes one special virtual reaction r_{growth} that represents cell growth, and the cell growth flux is represented by v_{growth}. S is a stoichiometry matrix, where $S_{ij} = k$ means that r_j produces k of m_i per unit time. If k is a negative number, then m_i is consumed. Let $V = \{v_1, \ldots, v_b\}$ be a set of reaction rates per unit time (flux) of R. Let $L = \{l_1, \ldots, l_b\}$ and $U = \{u_1, \ldots, u_b\}$ be the sets of the lower and upper bounds for V, respectively.

$C_1 = (M, R, S, L, U)$ is called the **metabolic network part** of C. v_{growth} is called the **growth rate** (**GR**). In FBA using C_1, GR is maximized by the following linear programming (LP):

maximize

v_{growth}

such that

$\Sigma_j S_{ij} v_j = 0$ for all i

$l_j \leq v_j \leq u_j$ for all j

$i = \{1, \ldots, a\}$, $j = \{1, \ldots, b\}$

If the ith column of S has only one non-zero element; in other words, r_i connects to only one metabolite, then r_i is called an **external reaction**, and is considered to be connected to the external environment. Reactions that are not exchange reactions are called **internal reactions**. The flux of the external reaction producing the target metabolite under the condition that cell growth is maximized is called the **production rate** (**PR**).

In contrast, $C_2 = (G, F, P)$ is called the **GPR network part** of C, and

$$p_i = f_i(g_{i,1}, \ldots, g_{i,k_i}), \quad \text{where} \quad p, g \in \{0, 1\} \text{ and } 1 \leq k_i \leq c.$$

If $p_i = 0$, then l_i and u_i are forced to be 0. In other words,

$$\begin{cases} v_i = 0 & \text{when} \quad p_i = 0, \\ l_i \leq v_i \leq u_i & \text{when} \quad p_i = 1 \end{cases}$$

hold.

The main problem of this study is formalized as follows.

Given

C, r_{target}, $PR_{threshold}$, $GR_{threshold}$

Find

$D \subset G$

such that minimizes

v_{target}

such that maximizes

v_{growth}

such that

$\Sigma_j S_{ij} v_j = 0$ for all i

$$\begin{cases} v_j = 0 \ \textit{if} \ p_j = 0 \\ l_j \leq v_j \leq u_j, \ \textit{otherwise} \end{cases}$$

$p_j = f_j(g_{j,1}, \ldots, g_{j,k_j})$

$$\begin{cases} g{=}0 \ \text{ if } \ \text{g} \in D \\ g = 1, \textit{otherwise} \end{cases}$$

$v_{target} \geq PR_{threshold}$

$v_{growth} \geq GR_{threshold}$

$p, g \in \{0, 1\}$

$$i = \{1, \ldots, a\},\ j = \{1, \ldots, b\}$$

Example for problem setting

Sub figure 2A shows a small toy example of the constraint-based model, where $M = \{m_1, \ldots, m_4\}$, $R = \{r_1, \ldots, r_7\}$, and $S = \begin{pmatrix} 1 & -1 & -1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & -1 \end{pmatrix}$.

Because $[\alpha, \beta]$ attached to r_i means that $\alpha \leq v_i \leq \beta$, $L = \{l_1, \ldots, l_7\}$ and $U = \{u_1, \ldots, u_7\}$ are as follows; $l_1, \ldots, l_7 = 0$, $u_1, \ldots, u_3 = 10$, $u_4, u_5 = 5$, $u_6, u_7 = 10$.

For C_2, it is given that $G = \{g_1, \ldots, g_5\}$, $F = \{f_1, \ldots, f_7\}$ and

$$\begin{aligned} p_1 &= f_1 = \phi, \\ p_2 &= f_2 = g1 \wedge g2 \wedge g3, \\ p_3 &= f_3 = \phi, \\ p_4 &= f_4 = g2 \wedge g5, \\ p_5 &= f_5 = (g3 \vee g4) \wedge g5, \\ p_6 &= f_6 = \phi, \\ p_7 &= f_7 = \phi. \end{aligned}$$

Note that f represents a Boolean function, whereas p takes either 0 or 1. $p_i = f_i = \phi$ means that r_i cannot be repressed via gene deletions.

In the original state, when GR (v_6) is maximized, all fluxes from r_1 flow through r_2 to r_6. Therefore, $v_1 = v_2 = v_6 = 10$ and $v_3 = v_4 = v_5 = v_7 = 0$ are obtained as shown in the second row of Sub figure 2B.

If g_1 is deleted, then $p_2 = g_1 \wedge g_2 \wedge g_3 = 0$ since $g_1 = 0$. Therefore, r_2 does not work and v_2 is forced to be zero. Similarly, r_3 does not work and v_3 is forced to be zero because $p_3 = g_1 = 0$ holds. Therefore, when GR is maximized, fluxes from r_1 cannot reach r_6, and GR becomes 0. In the optimistic case for PR, $v_1 = v_5 = v_7 = 5$ is obtained, but no flux flows in the pessimistic case, as shown in the third and fourth rows of Sub figure 2B, respectively. To ensure the growth coupling, we need to evaluate the pessimistic case for PR, and the maximized GR must exceed the minimum required value. Therefore, we consider that growth coupling cannot be achieved by deleting g_1. When g_2 is deleted, similar results are obtained because neither r_2 nor r_4 works.

If g_3 is deleted, r_2 does not work but the other reactions can work. Therefore, the maximum GR is five because $0 \leq v_4 \leq 5$. In the optimistic case, the flux from r_1 flows to r_7 via r_5 in addition to via r_3 and r_4. In this case, GR=5 and PR=10 is obtained. However, in the pessimistic case, GR=PR=5 were obtained as shown in the seventh and eighth rows of Sub figure 2B, respectively.

If g_4 is deleted, $p_i = 1$ for all i. Therefore, $v_1 = v_2 = v_6 = 10$ and $v_3 = v_4 = v_5 = v_7 = 0$ are obtained when GR is maximized. If g_5 is deleted, neither r_4 nor r_5 works since $p_4 = p_5 = 0$. However, a similar result is obtained because r_2 works as shown in the ninth and tenth rows of Sub figure 2B, respectively.

Suppose that GRLB=PRLB=1; that is, the minimum required GR and PR are 1. Then, deleting g_3 achieves growth coupling because GR=PR=5 is obtained even for the pessimistic case and GR≥GRLB and PR≥PRLB are satisfied. In this example, growth coupling can be achieved by deleting one gene g_3. However, in practice, it may be necessary to examine all genes on and off, which results in a combinatorial explosion.

Pseudo code

The pseudo code of gDel_minRN is as follows.

Procedure **gDel_minRN**$(model, targetMet, PRLB, GRLB, maxloop)$

/*Calculating the theoretical maximum production rate.*/

$TMPR =$**max** v_{target}

s.t. $\Sigma_j \ S_{i,j} \cdot v_j = 0$ for all $1 \leq i \leq a$

$LB_j \leq v_j \leq UB_j$ for all $1 \leq j \leq b$

if $TMPR < PRLB$

return "no solution"

/*Calculating the theoretical maximum growth rate.*/

$TMGR =$**max** v_{growth}

s.t. $\Sigma_j \ S_{i,j} \cdot v_j = 0$ for all $1 \leq i \leq a$

$LB_j \leq v_j \leq UB_j$ for all $1 \leq j \leq b$

/* Finding a gene deletion strategy candidate.*/

$prohibited\_list = \phi$, $loop = 1$

while $loop \leq maxloop$

max $v_{growth} + TMGR \cdot KO$ /*first maximize #repressed reactions.*/

s.t. $\Sigma_j S_{i,j} \cdot v_j = 0$

$$\begin{cases} v_j = 0 \ if \ p_j = 0 \\ l_j \leq v_j \leq u_j, \ otherwise \end{cases}$$

$p_j = f_j(g_{j,1}, \ldots, g_{j,k_j})$

KO: the number of repressed reactions ($p_j = 0$).

$$\begin{cases} g = 0 \ \text{ if } \ g \in D \ /^*D \text{ is flexible.}^*/ \\ g = 1, otherwise \end{cases}$$

$D \notin prohibited\_list$

GRLB $\leq v_{growth}$

PRLD $\leq v_{target}$

$D_{candidate} = D$

/*Checking whether growth coupling is achieved by $D_{candidate}$.*/

min

v_{target}

such that max

v_{growth}

such that

$\Sigma_j S_{ij} v_j = 0$ for all i

$$\begin{cases} v_j = 0 \ if \ p_j = 0 \\ l_j \leq v_j \leq u_j, \ otherwise \end{cases}$$

$p_j = f_j(g_{j,1}, \ldots, g_{j,k_j})$

$$\begin{cases} g = 0 \ \text{ if } \ g \in D_{candidate} \ /^*D_{candidate} \text{ is fixed.}^*/ \\ g = 1, otherwise \end{cases}$$

if $v_{target} \geq$PRLB and $v_{growth} \geq$GRLB **then**

return $D_{candidate}$, v_{target}, v_{growth}

else

$prohibited\_list = prohibited\_list \cup D$

$loop = loop + 1$

List of abbreviations

Gene-Protein-Reaction (GPR)

Flux Balance Analysis (FBA)

Growth Rate (GR)

Production Rate (PR)

Mixed-Integer Linear Programming (MILP)

Theoretical Maximum Production Rate (TMPR)

Theoretical Maximum Growth Rate (TMGR)

Acyl-Carrier-Protein (ACP)

Nicotinamide Adenine Dinucleotide (NAD)

Nicotinamide Adenine Dinucleotide Phosphate (NADP)

Ubiquinone (UQ)

Adenosine Triphosphate (ATP)

Linear Programming (LP)

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the GitHub repository, https://github.com/taketam/gDel-minRN

Competing interests

The authors declare that they have no competing interests.

Funding

TT, AMF, YT, and TK were partially supported by grants from JSPS, KAKENHI #20H04242. No funding body played any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

TT designed this work, developed the algorithm, implemented the software, and conducted the computational experiments. AMF performed the analysis using the database. YT performed the visualization analysis. TK provided the biological interpretation of the experimental results. TT and TK wrote the manuscript. All authors have read the manuscript and approved it.

Acknowledgements

We would like to thank all the editors and reviewers involved in this paper.

Author details

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan. [2]Center for Biosystems Dynamics Research, RIKEN, Suita, Osaka, Japan. [3]Faculty of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, Japan. [4]Research Center for Thermotolerant Microbial Resources (RCTMR), Yamaguchi University, Yoshida, Yamaguchi, Japan. [5]Graduate School of Sciences and Technology for Innovation, Yamaguchi University, Yoshida, Yamaguchi, Japan.

References

1. Burgard, A.P., Pharkya, P., Maranas, C.D.: Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnology and bioengineering **84**(6), 647–657 (2003)
2. Pharkya, P., Burgard, A.P., Maranas, C.D.: Optstrain: a computational framework for redesign of microbial production systems. Genome research **14**(11), 2367–2376 (2004)
3. Pharkya, P., Maranas, C.D.: An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. Metabolic engineering **8**(1), 1–13 (2006)
4. Patil, K.R., Rocha, I., Förster, J., Nielsen, J.: Evolutionary programming as a platform for in silico metabolic engineering. BMC bioinformatics **6**(1), 308 (2005)
5. Ranganathan, S., Suthers, P.F., Maranas, C.D.: Optforce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. PLoS Comput Biol **6**(4), 1000744 (2010)
6. Rocha, I., Maia, P., Evangelista, P., Vilaça, P., Soares, S., Pinto, J.P., Nielsen, J., Patil, K.R., Ferreira, E.C., Rocha, M.: Optflux: an open-source software platform for in silico metabolic engineering. BMC systems biology **4**(1), 1–12 (2010)
7. Toya, Y., Shimizu, H.: Flux analysis and metabolomics for systematic metabolic engineering of microorganisms. Biotechnology advances **31**(6), 818–826 (2013)
8. Orth, J.D., Thiele, I., Palsson, B.Ø.: What is flux balance analysis? Nature biotechnology **28**(3), 245–248 (2010)
9. Lun, D.S., Rockwell, G., Guido, N.J., Baym, M., Kelner, J.A., Berger, B., Galagan, J.E., Church, G.M.: Large-scale identification of genetic design strategies using local search. molecular systems biology **5**(1), 296 (2009)
10. Rockwell, G., Guido, N.J., Church, G.M.: Redirector: designing cell factories by reconstructing the metabolic objective. PLoS Comput Biol **9**(1), 1002882 (2013)
11. Yang, L., Cluett, W.R., Mahadevan, R.: Emilio: a fast algorithm for genome-scale strain design. Metabolic engineering **13**(3), 272–281 (2011)
12. Egen, D., Lun, D.S.: Truncated branch and bound achieves efficient constraint-based genetic design. Bioinformatics **28**(12), 1619–1623 (2012)
13. Lewis, N.E., Hixson, K.K., Conrad, T.M., Lerman, J.A., Charusanti, P., Polpitiya, A.D., Adkins, J.N., Schramm, G., Purvine, S.O., Lopez-Ferrer, D., *et al.*: Omic data from evolved $e.\ coli$ are consistent with computed optimal growth from genome-scale models. Molecular systems biology **6**(1), 390 (2010)
14. Gu, D., Zhang, C., Zhou, S., Wei, L., Hua, Q.: Idealknock: a framework for efficiently identifying knockout strategies leading to targeted overproduction. Computational biology and chemistry **61**, 229–237 (2016)
15. Ohno, S., Shimizu, H., Furusawa, C.: Fastpros: screening of reaction knockout strategies for metabolic engineering. Bioinformatics **30**(7), 981–987 (2014)
16. Tamura, T.: Grid-based computational methods for the design of constraint-based parsimonious chemical reaction networks to simulate metabolite production: Gridprod. BMC bioinformatics **19**(1), 325 (2018)
17. von Kamp, A., Klamt, S.: Growth-coupled overproduction is feasible for almost all metabolites in five major production organisms. Nature communications **8**, 15956 (2017)
18. Machado, D., Herrgård, M.J., Rocha, I.: Stoichiometric representation of gene–protein–reaction associations leverages constraint-based analysis from reaction to gene-level phenotype prediction. PLoS computational biology **12**(10), 1005140 (2016)
19. Razaghi-Moghadam, Z., Nikoloski, Z.: Genereg: A constraint-based approach for design of feasible metabolic engineering strategies at the gene level. Bioinformatics (2020)
20. Acevedo-Rocha, C., Gronenberg, L., Mack, M., Commichau, F., Genee, H.: Microbial cell factories for the sustainable manufacturing of b vitamins. Curr. Opin. Biotechnol. **56**, 18–29 (2019)

21. Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S.N., Richelle, A., Heinken, A., Haraldsdóttir, H.S., Wachowiak, J., Keating, S.M., Vlasov, V., *et al.*: Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0. Nature protocols **14**(3), 639–702 (2019)
22. Monk, J.M., Lloyd, C.J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., *et al.*: iml1515, a knowledgebase that computes $escherichia\ coli$ traits. Nature biotechnology **35**(10), 904–908 (2017)
23. King, Z.A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N.E., Palsson, B.O.: Escher: a web application for building, sharing, and embedding data-rich visualizations of biological pathways. PLoS Comput Biol **11**(8), 1004321 (2015)
24. Kanehisa, M., Sato, Y.: Kegg mapper for inferring cellular functions from protein sequences. Protein Science **29**(1), 28–35 (2020)
25. Xiao, F., Wang, H., Shi, Z., Huang, Q., Huang, L., Lian, J., Cai, J., Xu, Z.: Multi-level metabolic engineering of *Pseudomonas mutabilis* atcc31014 for efficient production of biotin. Metab Eng **61**, 406–415 (2020)
26. Booth, I., Ferguson, G., Miller, S., Li, C., Gunasekera, B., Kinghorn, S.: Bacterial production of methylglyoxal: a survival strategy or death by misadventure. Biochem Soc Trans **31**(Pt 6), 1406–1408 (2003)
27. Niu, W., Kramer, L., Mueller, J., Liu, K., Guo, J.: Metabolic engineering of *Escherichia coli* for the de novo stereospecific biosynthesis of 1,2-propanediol through lactic acid. Metab Eng Commun **8**, 00082 (2019)
28. Apaolaza, I., Valcarcel, L.V., Planes, F.J.: gmcs: fast computation of genetic minimal cut sets in large networks. Bioinformatics **35**(3), 535–537 (2019)
29. Kim, J., Reed, J.L., Maravelias, C.T.: Large-scale bi-level strain design approaches and mixed-integer programming solution techniques. PLoS One **6**(9), 24162 (2011)
30. Tepper, N., Shlomi, T.: Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. Bioinformatics **26**(4), 536–543 (2010)

Figures

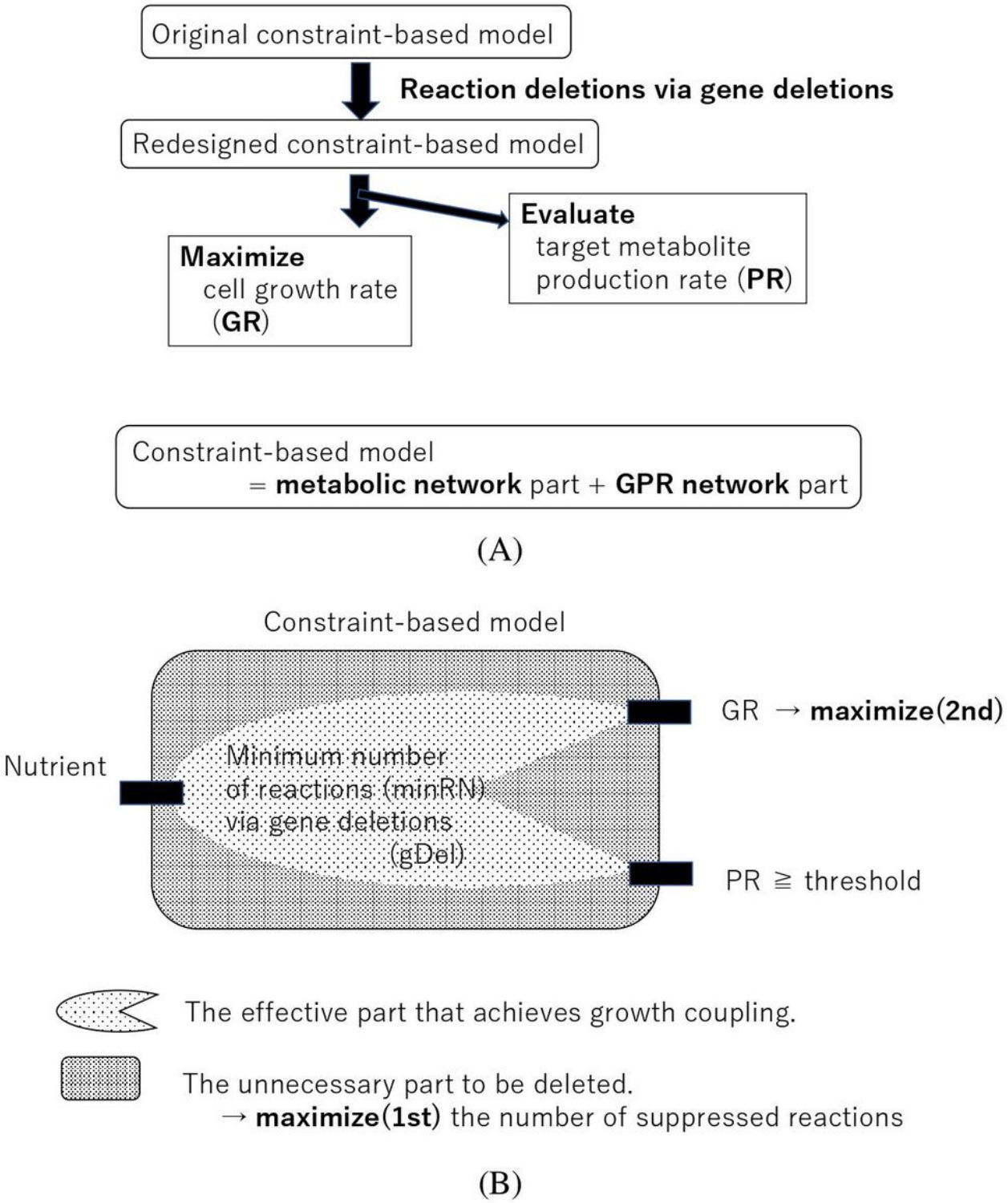
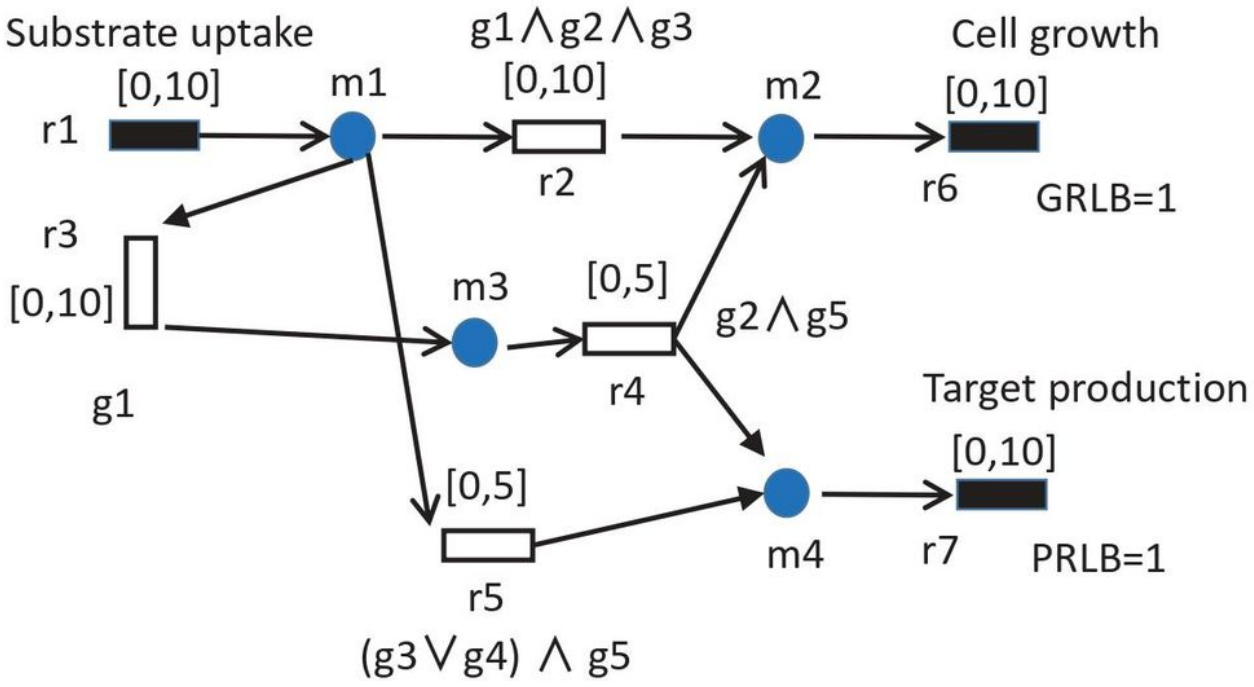


Figure 1

(A) Problem setting of this study. PR of the target metabolite is evaluated when the GR is maximized. (B) The idea of gDel minRN algorithm. The maximum number of reactions are repressed via gene deletions for the growth coupling.



(A)

Gene KO		v_1	v_2	v_3	v_4	v_5	v_6	v_7
none	both	10	10	0	0	0	10	0
g1	best	5	0	0	0	5	0	5
	worst	0	0	0	0	0	0	0
g2	best	5	0	0	0	5	0	5
	worst	0	0	0	0	0	0	0
g3	best	10	0	5	5	5	5	10
	worst	5	0	5	5	0	5	5
g4	both	10	10	0	0	0	10	0
g5	both	10	10	0	0	0	10	0
g1, g2	both	0	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(B)

Figure 2

(A) A toy example of the constraint-based model. Circles and rectangles represent metabolites and reactions, respectively. Black and white rectangles are external and internal reactions. r1, r6, and r7 are the substrate uptake, cell growth, and target metabolite production reactions. [α, β] represents the lower and upper bounds of the reaction rates. (B) The optimistic and pessimistic flux distributions from the

viewpoints of PR for each gene deletion strategy when GR is maximized. Deleting g3 achieves the growth coupling since PR≥PRLB and GR≥GRLB are satisfied even for the pessimistic case of PR.

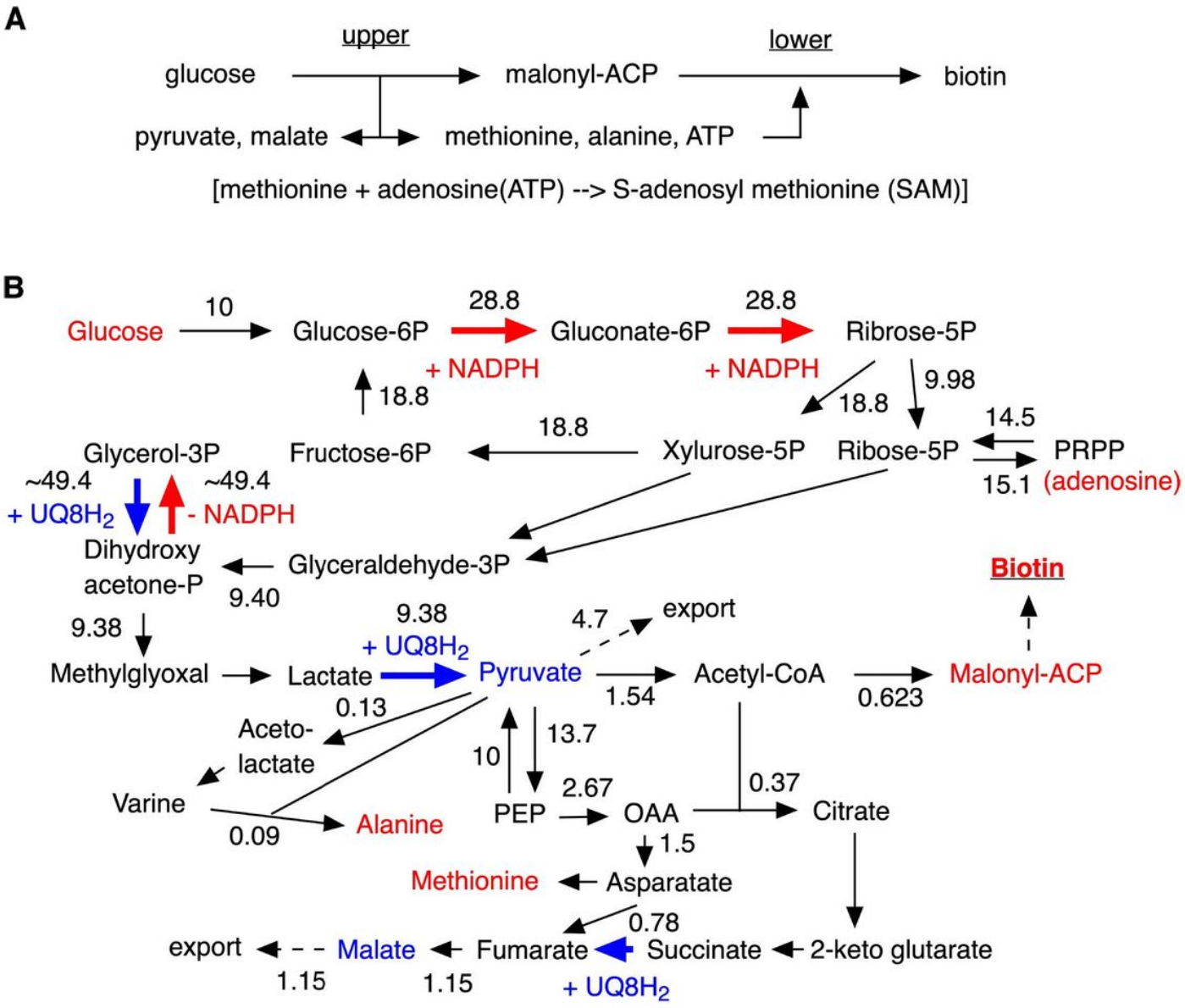


Figure 3

The constructed pathway for biotin production. (A) Overview of the biotin synthesis pathway from iML1515 classified into two pathways as upper and lower pathway. (B) Precise flow of upper pathway, from glucose to malonyl-ACP. The number indicated with each arrows shows the flux value of each reaction. The abbreviations are as follows; NADPH, Nicotinamide adenine dinucleotide phosphate reduced form; UQ8, ubiquinone-8; ACP, acyl carrier protein; PEP, phosphoenolpyruvate; OAA, oxaloacetate; PRPP, Phosphoribosyl diphosphate.